RATIO STATISTICS

František Vávra, Tomáš Pavelka, Blanka Šedivá, Kateřina Vokáčová, Patrice Marek, Martina Neumanová

Keywords: Ratio statistics, tolerance bounds, point estimates, stochastic ordering

Klíčová slova: Poměrové statistiky, toleranční meze, bodové odhady, stochastické srovnávání

Abstract: The results of independent experiments used for testing of certain method or certain phenomenon are often represented by samples (r_1, s_1), ..., (r_n, s_n), where $r_i, s_i \geq 0; \forall i = 1, \ldots, n; \sum_{i=1}^{n} r_i > 0$. The main goal of our contribution is to analyse the random variable $X = \frac{R}{R+S}$. The analysis is concerned with tolerance bounds of statistic $X_n = \frac{\sum_{i=1}^{n} r_i}{\sum_{i=1}^{n} r_i + \sum_{i=1}^{n} s_i}$ and introduces its asymptotic model as well. Furthermore, two particular cases of n are examined. In the first case, n is assumed to be non-random but changeable (e.g. comparison of automatic speech recognition methods). In the other case, n is assumed to be fixed (e.g. calculation of state unemployment rate from all district unemployment rates).

Abstrakt: Předložená práce se zabývá studiem tolerančních mezí statistiky $X_n = \frac{\sum_{i=1}^{n} r_i}{\sum_{i=1}^{n} r_i + \sum_{i=1}^{n} s_i}$, kde (r_1, s_1), ..., (r_n, s_n) jsou pozorování, které popisují výsledky nezávislých experimentů testujících nějakou metodu nebo jev, $r_i, s_i \geq 0; \forall i = 1, \ldots, n$; $\sum_{i=1}^{n} r_i > 0$. O statistice se předpokládá, že měřená veličina, kterou představuje, má tvar $X = \frac{R}{R+S}$. V práci je prezentován asymptotický model této statistiky. Jsou analyzovány dva dílčí případy. V prvním je n nenáhodné ale proměnné (např. testy úspěšnosti metod pro rozpoznávání mluvených slov). V druhém případě je n pevně dané a neměnné (míra nezaměstnanosti státu počítaná z nezaměstnanosti všech okresů).

1 Motivation and Model

The observations of the results of independent experiments measuring phenomenon or testing a method are often represented as (r_1, s_1), ..., (r_n, s_n), where $r_i, s_i \geq 0; \forall i = 1, \ldots, n; \sum_{i=1}^{n} r_i > 0$. The expected value, the variance and the covariance of each variable are known. $\forall i = 1, \ldots, n$, $E\{r_i\} = e_{ir}$, $E\{s_i\} = e_{is}$, $\sigma^2(r_i) = \sigma_{ir}^2$, $\sigma^2(s_i) = \sigma_{is}^2$ and $\forall i = 1, \ldots, n; j = 1, \ldots, n$; $E\{(s_i - e_{is})(r_j - e_{jr})\}$. The quality of the tested method or the rate of

certain phenomenon is usually measured by the following criterion

$$X_n = \frac{\sum_{i=1}^{n} r_i}{\sum_{i=1}^{n} r_i + \sum_{i=1}^{n} s_i}.$$

This criterion is used in many tasks. One example of such application, which is demonstrated in our contribution, is the testing of automatic speech recognition methods, where:

r_i represents the number of recognised words in i-th sentence of the length N_i,

s_i represents the number of incorrectly recognised words in i-th sentence of the length N_i,

N_i represents the length of the sentence in number of words,

n represents the number of sentences on which the method is tested.

The number n is changeable, i.e. can be different for different tests. Another example of application is the measurement of the unemployment rate of the whole country, which uses the numbers of the unemployed in all particular districts for its evaluation. In this case,

r_i represents the number of the registered unemployed in i-th district,

s_i represents the number of the employed in i-th district,

n represents the number of districts which is fixed, i.e. it is the same for all the tests performed.

2 Probability description

The criterion

$$X_n = \frac{\sum_{i=1}^{n} r_i}{\sum_{i=1}^{n} r_i + \sum_{i=1}^{n} s_i}$$

is the random variable, which is computed from the results of the experiments. The criterion could be transformed to

$$X_n = \frac{1}{1 + \frac{\sum_{i=1}^{n} s_i}{\sum_{i=1}^{n} r_i}} = \frac{1}{1 + Y_n}.$$

Let us assume the distribution function of the variable $Y_n = \frac{\sum_{i=1}^{n} s_i}{\sum_{i=1}^{n} r_i}$, $F_{Y_n}(x)$ is known. Then the distribution function

$$F_{X_n}(x) = P\{X_n < x\} = P\left\{\tfrac{1}{1+Y_n} < x\right\}$$

could be expressed as follows

$$F_{X_n}(x) = 1 - F_{Y_n}(x)\left(\tfrac{1}{x} - 1\right) \Leftrightarrow 0 < x < 1;\ F_{X_n}(x) = 0 \Leftrightarrow x \leq 0;$$
$$F_{X_n}(x) = 1 \Leftrightarrow x \geq 1.^{1}$$

Distribution function of variable Y_n for $x > 0$

$$F_{Y_n}(x) = P\{Y_n < x\} = P\left\{\frac{\sum_{i=1}^{n} s_i}{\sum_{i=1}^{n} r_i} < x\right\} = P\left\{\sum_{i=1}^{n} s_i - x\sum_{i=1}^{n} r_i < 0\right\} = (A)$$

Further on, the following notation will be used $e_r = \frac{1}{n}\sum_{i=1}^{n} e_{ir}$ and $e_s = \frac{1}{n}\sum_{i=1}^{n} e_{is}$. The random variable $\sum_{i=1}^{n} s_i - x\sum_{i=1}^{n} r_i$ has the expected value of

$$E\left\{\sum_{i=1}^{n} s_i - x\sum_{i=1}^{n} r_i\right\} = n(e_s - xe_r).$$

The variance of this variable has the following form

$$\sigma^2\left\{\sum_{i=1}^{n} s_i - x\sum_{i=1}^{n} r_i\right\} = n\left[\sigma_s^2 + x^2\sigma_r^2 - 2x\rho_{s,r}\sigma_s\sigma_r\right],$$

where
$$\sigma_r^2 = \tfrac{1}{n}\sum_{i=1}^{n} E\left\{(r_i - e_r)^2\right\},$$
$$\sigma_s^2 = \tfrac{1}{n}\sum_{i=1}^{n} E\left\{(s_i - e_s)^2\right\},$$
$$\rho_{s,r}\sigma_s\sigma_r = \tfrac{1}{n}\sum_{i=1}^{n} E\left\{(s_i - e_s)(r_i - e_r)\right\}$$

and x is a fixed known value for which the distribution function is computed. We can continue with the derivation of the distribution function $F_{Y_n}(x)$:

$$\begin{aligned}(A) &= P\left\{\sum_{i=1}^{n} s_i - x\sum_{i=1}^{n} r_i < 0\right\} = \\ &= P\left\{\sum_{i=1}^{n} s_i - x\sum_{i=1}^{n} r_i - n(e_s - xe_r) < -n(e_s - xe_r)\right\} = \\ &= P\left\{\frac{\sum_{i=1}^{n} s_i - x\sum_{i=1}^{n} r_i - n(e_s - xe_r)}{\sqrt{n(\sigma_s^2 + x^2\sigma_r^2 - 2x\rho_{s,r}\sigma_s\sigma_r)}} < \frac{-n(e_s - xe_r)}{\sqrt{n(\sigma_s^2 + x^2\sigma_r^2 - 2x\rho_{s,r}\sigma_s\sigma_r)}}\right\} = (B)\end{aligned}$$

[1] Any possible discontinuities of Y are not discussed in our contribution. The eventual occurrence of the discontinuity could be neglected in the supposed applications.

The left side of the inequality $\frac{\sum_{i=1}^{n} s_i - x\sum_{i=1}^{n} r_i - n(e_s - xe_r)}{\sqrt{n(\sigma_s^2 + x^2\sigma_r^2 - 2x\rho_{s,r}\sigma_s\sigma_r)}}$ is a centralized and standardized random variable whose distribution function for increasing n converges to the distribution function of the standard normal distribution

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{z^2}{2}} dz.$$

(The proofs with different variants of additional assumptions are described in e.g. [1] p. 374 - 382.) Therefore, for a large enough n the following is valid:

$$(B) = \Phi\left(-\sqrt{n}\frac{e_s - xe_r}{\sqrt{\sigma_s^2 + x^2\sigma_r^2 - 2x\rho_{s,r}\sigma_s\sigma_r}}\right).$$

Summing up: asymptotically for $x > 0$,

$$F_{Y_n}(x) = P\{Y_n < x\} = \Phi\left(-\sqrt{n}\frac{e_s - xe_r}{\sqrt{\sigma_s^2 + x^2\sigma_r^2 - 2x\rho_{s,r}\sigma_s\sigma_r}}\right).$$

Distribution function of variable X_n
The distribution function of the criterion X_n can be described as follows:

$$F_{X_n}(x) = 1 - \Phi\left(-\frac{\sqrt{n}\cdot\left(e_s - (\frac{1}{x} - 1)e_r\right)}{\sqrt{(\sigma_s^2 + (\frac{1}{x} - 1)^2\sigma_r^2 - 2(\frac{1}{x} - 1)\rho_{s,r}\sigma_s\sigma_r)}}\right) \Leftrightarrow 0 < x < 1.$$

Such expression is quite complicated, nevertheless easy to evaluate.

3 Point estimate of the measure of location

The most representative measure of location for such type of distribution is the median, which is the solution of the equation $F_{X_n}(x_{med}) = \frac{1}{2}$. Thus

$$e_s = \left(\frac{1}{x_{med}} - 1\right)\cdot e_r \Rightarrow x_{med} = \frac{e_r}{e_r + e_s}.$$

4 Tolerance interval

The goal of this part is to find bounds $x_L < x_U$ such that

$$P(x_L \leq X_n \leq x_U) = 1 - \alpha,$$

where α is the confidence level. In other words, we want to find the percentile range that represents a specified proportion of the population. The relation $P(x_L \leq X_n \leq x_U) = 1 - \alpha$ does not uniquely determine the bounds. Therefore, α_1, α_2 such that

$$\alpha = \alpha_1 + \alpha_2; \quad 0 < \alpha, \alpha 1, \alpha 2 < 1$$

are chosen to satisfy $P(x_U < X_n) = \alpha_1 \Leftrightarrow P(X_n < x_U) = 1 - \alpha_1$ and $P(X_n < x_L) = \alpha_2$. The particular bound determination is realized by solving the equations for the distribution function $F_{X_n}(x_U) = 1 - \alpha_1, F_{X_n}(x_L) = \alpha_2$.

Determination of the lower bound x_L
x_L is the solution of the equation

$$1 - \Phi\left(-\sqrt{n} \frac{e_s - \frac{1}{x_L} - 1 \; e_r}{\sqrt{\sigma_s^2 + \frac{1}{x_L} - 1 \;^2 \sigma_r^2 - 2 \; \frac{1}{x_L} - 1 \; \rho_{s,r}\sigma_s\sigma_r}} \right) = \alpha_2, \text{ thus}$$

$$\Phi^{-1}(1 - \alpha_2) = \left(-\sqrt{n} \frac{e_s - \frac{1}{x_L} - 1 \; e_r}{\sqrt{\sigma_s^2 + \frac{1}{x_L} - 1 \;^2 \sigma_r^2 - 2 \; \frac{1}{x_L} - 1 \; \rho_{s,r}\sigma_s\sigma_r}} \right).$$

If the notation $z = \frac{1}{x_L} - 1$ is used, the equation has the form of:

$$\Phi^{-1}(1 - \alpha_2) = \left(-\sqrt{n} \frac{e_s - z e_r}{\sqrt{\sigma_s^2 + z^2\sigma_r^2 - 2z\rho_{s,r}\sigma_s\sigma_r}} \right) \Rightarrow$$
$$\Phi^{-1}(1 - \alpha_2)\sqrt{\sigma_s^2 + z^2\sigma_r^2 - 2z\rho_{s,r}\sigma_s\sigma_r} = -\sqrt{n}(e_s - z e_r)$$

The next equation modification is squaring. By this step, the solution is represented by one of the roots of the equation

$$\left[\Phi^{-1}(1 - \alpha_2)\right]^2 \left(\sigma_s^2 + z^2\sigma_r^2 - 2z\rho_{s,r}\sigma_s\sigma_r\right) = n\left(e_s - z e_r\right)^2,$$

which will be identified by backward testing of the validity of these roots on the equation before the squaring. Finally, this letter notation was used to express the equation in a more transparent form:

$$Az^2 + 2zB + C = 0,$$

where $A = (a\sigma_r^2 - e_r^2)$,
$B = (e_s e_r - \rho_{s,r} a \sigma_r \sigma_s)$,
$C = a\sigma_s^2 - e_s^2$ and
$a = \frac{\left[\Phi^{-1}(1-\alpha_2)\right]^2}{n}$.

Two roots $z_{1,2} = \frac{-B \pm \sqrt{B^2 - AC}}{A}$ are evaluated by solving this quadratic equation. As mentioned before, the identification of the valid solution is arranged by checking whether the particular root satisfies the following equation

$$\frac{\Phi^{-1}(1-\alpha_2)}{\sqrt{n}} \sqrt{\sigma_s^2 + z^2\sigma_r^2 - 2z\rho_{s,r}\sigma_s\sigma_r} + (e_s - z e_r) = 0.$$

Only one root is valid. Let us denote the solution $z\star$. Using the transformation $z\star = \frac{1}{x_L} - 1$, the solution $x_L = \frac{1}{1+z\star}$ is obtained.

Determination of the upper bound x_U
The determination of the x_U results from solving the equation

$$\Phi\left(-\sqrt{n}\frac{e_s-\left(\frac{1}{x_U}-1\right)e_r}{\sqrt{\sigma_s^2+\left(\frac{1}{x_U}-1\right)^2\sigma_r^2-2\left(\frac{1}{x_U}-1\right)\rho_{s,r}\sigma_s\sigma_r}}\right)=\alpha_1.$$

The calculation is almost identical to the previous calculation of x_L. The only differences are

$$a=\frac{\left[\Phi^{-1}(\alpha_1)\right]^2}{n}$$

and that the equation for testing the validity of the roots is

$$\frac{\Phi^{-1}(\alpha_1)}{\sqrt{n}}\sqrt{\sigma_s^2+z^2\sigma_r^2-2z\rho_{s,r}\sigma_s\sigma_r}+(e_s-ze_r)=0.$$

5 Acceptability of asymptotic representation

The formula of the asymptotic version of the distribution function $F_{X_n}(x)$ does not necessarily need to have the characteristics of the distribution function, and it usually has not. Furthermore, X_n is the random variable defined on the interval $(0,1\rangle$. The formula of $F_{X_n}(x)$ exceeds this interval by not having the functional value equal to zero or one in the boundary elements of the interval. These circumstances could be used for testing the acceptability of such asymptotic representation. For these purposes, following assumptions are made:

1. The function $F_{X_n}(x)$ should be a non-decreasing function in the interval $(0,1\rangle$.
2. The value of $F_{X_n}(x)$ for x in the "zero value" should be non-negative and smaller than a known small positive number β_0. The value of $F_{X_n}(x)$ for x in the point $x=1$ should not exceed one and should be greater than $1-\beta_1$, where β_1 is positive and sufficiently small.

The first assumption
The first assumption is fulfilled if $\frac{d}{dx}F_{X_n}(x)=f_{X_n}(x)$ is non-negative in the interval $(0,1\rangle$. The artificial variable $z=\frac{1}{x}-1$ is used again for the sake of transparency of the expressions mentioned below.

$$F_{X_n}(z)=1-\Phi\left(-\sqrt{n}\frac{e_s-ze_r}{\sqrt{\sigma_s^2+z^2\sigma_r^2-2z\rho_{s,r}\sigma_s\sigma_r}}\right);\ 0<z<+\infty \text{ and } \frac{dz}{dx}=-\frac{1}{x^2}.$$

Then

$$f_{X_n}(x)=\frac{d}{dx}F_{X_n}(x)=\frac{d}{dz}F_{X_n}(z)\frac{dz}{dx}=$$
$$\frac{d}{dz}\left[1-\Phi\left(-\sqrt{n}\frac{e_s-ze_r}{\sqrt{\sigma_s^2+z^2\sigma_r^2-2z\rho_{s,r}\sigma_s\sigma_r}}\right)\right]\frac{dz}{dx}.$$

Taking into account the negative sign of $\frac{dz}{dx}$, the sign of $f_{X_n}(x)$ is decisive.

$$\frac{d}{dz}\left[1-\Phi\left(-\sqrt{n}\frac{e_s-ze_r}{\sqrt{\sigma_s^2+z^2\sigma_r^2-2z\rho_{s,r}\sigma_s\sigma_r}}\right)\right]=$$
$$-\sqrt{n}\varphi\left(-\sqrt{n}\frac{e_s-ze_r}{\sqrt{\sigma_s^2+z^2\sigma_r^2-2z\rho_{s,r}\sigma_s\sigma_r}}\right)\left(\frac{e_r\left(\sigma_s^2-z\rho_{s,r}\sigma_s\sigma_r\right)+e_s\left(z\sigma_r^2-\rho_{s,r}\sigma_s\sigma_r\right)}{\left(\sigma_s^2+z^2\sigma_r^2-2z\rho_{s,r}\sigma_s\sigma_r\right)^{\frac{3}{2}}}\right)$$

The condition $f_{X_n}(x)\geq 0$ is fulfilled for such $z=\frac{1}{x}-1$, where

$$e_r\left(\sigma_s^2-z\rho_{s,r}\sigma_s\sigma_r\right)+e_s\left(z\sigma_r^2-\rho_{s,r}\sigma_s\sigma_r\right)\geq 0.$$

The first step of the test is to find the intercept point of the monotonic function, i.e. the solution of the equation

$$e_r\left(\sigma_s^2-z\rho_{s,r}\sigma_s\sigma_r\right)+e_s\left(z\sigma_r^2-\rho_{s,r}\sigma_s\sigma_r\right)=0.$$

Thus

$$z=\frac{\sigma_s}{\sigma_r}\frac{\rho_{s,r}e_s\sigma_r-e_r\sigma_s}{e_s\sigma_r-\rho_{s,r}e_r\sigma_s}.$$

The other step is to check if the intercept point x is out of the interval $(0,1\rangle$. This is assured by $\frac{\rho_{s,r}e_s\sigma_r-e_r\sigma_s}{e_s\sigma_r-\rho_{s,r}e_r\sigma_s}\leq 0$, because $x=\frac{1}{1+z}$. The next step is the verification of the inequality in the median point.

$$x_{med}=\frac{e_r}{e_r+e_s}\Rightarrow z_{med}=\frac{e_s}{e_r},$$

hence

$$e_r\left(\sigma_s^2-z_{med}\rho_{s,r}\sigma_s\sigma_r\right)+e_s\left(z_{med}\sigma_r^2-\rho_{s,r}\sigma_s\sigma_r\right)\geq\frac{\left(e_r\sigma_s-e_s\sigma_r\right)^2}{e_r}\geq 0.$$

For this reason, the condition $\frac{\rho_{s,r}e_s\sigma_r-e_r\sigma_s}{e_s\sigma_r-\rho_{s,r}e_r\sigma_s}\leq 0$ represents the necessary and sufficient condition of $f_{X_n}(x)\geq 0$; $0<x\leq 1$. If this condition is satisfied, the asymptotic approximation will be considered correct. The correctness is important for the "small n".

Note: The mentioned correctness condition is not dependent on the number of observations.

The second assumption

When the numbers $0<\beta_0,\beta_1,\beta=\beta_0+\beta_1<1$ are set, the asymptotic approximation could be considered as β-acceptable, if the inequalities

$$1\geq F_{X_n}(1)\geq 1-\beta_1 \text{ and } 0\leq\lim_{x\to 0}F_{X_n}(x)\leq\beta_0$$

are fulfilled. It means that $F_{X_n}(1)=1-\Phi\left(-\sqrt{n}\frac{e_s}{\sigma_s}\right)$, so

$$1-\Phi\left(-\sqrt{n}\frac{e_s}{\sigma_s}\right)\geq 1-\beta_1\Rightarrow n\geq\frac{\sigma_s^2}{e_s^2}\left(\Phi^{-1}(\beta_1)\right)^2.$$

The other $\lim_{x\to 0_+}F_{X_n}(x)=1-\Phi(\sqrt{n}\frac{e_s}{\sigma_r})$, and so

$$1 - \Phi(\sqrt{n}\frac{e_s}{e_r}) \leq \beta_0 \Rightarrow n \geq \frac{\sigma_r^2}{e_r^2}(\Phi^{-1}(1-\beta_0))^2.$$

Summing up: The asymptotic approximation could be considered as $\beta = \beta_0 + \beta_1$ acceptable if and only if $n \geq max\left\{\frac{\sigma_s^2}{e_s^2}\left(\Phi^{-1}(\beta_1)\right)^2; \frac{\sigma_r^2}{e_r^2}\left(\Phi^{-1}(1-\beta_0)\right)^2\right\}$.

Note: The condition of β-acceptability is dependent on the observation number n.

6 Conclusion

The mentioned methods can be applied in the cases where the measurements are in the form of a ratio criterion. The ratio criterion takes the values from the interval $\langle 0; 1\rangle$ for the purposes of our contribution. The extension of this interval by "one" could be arranged just by a technical procedure. Nevertheless, it is necessary to solve a few problems with the correctness and acceptability of the asymptotic approximation. Further research can focus on the stochastic comparison.

References

[1] Rényi A. (1972) : *Teorie pravděpodobnosti*. ACADEMIA, Paris

Acknowledgement: The research has been supported by :

MSM4977751301 research plan - Continuous and Discrete Mathematical Structures and Development of Corresponding Methods of their Study,

MPO 2A - 2TP1/051 grant - Improving Reliability and Safety of Electrical Networks,

NPV II: 2C06009 Cot-Sewing grant - Complex knowledge base tools for natural language communication with the semantic web.

Adresa: ZČU FAV, KMA, Univerzitní 22, 306 14 Plzeň

E-mail: `vavra@kma.zcu.cz`